**CLAIMS:**

What is claimed is:

1. A method of selecting data sets for use with a predictive algorithm, comprising:

generating a first distribution of a training data set;

generating a second distribution of a testing data set;

comparing the first distribution and the second distribution to identify a discrepancy between the first distribution and the second distribution; and

modifying selection of entries in one or more of the training data set and the testing data set based on the discrepancy between the first distribution and the second distribution.

2. The method of claim 1, wherein the first distribution and the second distribution are distributions of drive time from a customer geographical location to a commercial establishment geographical location.

3. The method of claim 1, wherein the first distribution and the second distribution are distributions of distance between a customer geographical location and a commercial establishment geograhical location.

4. The method of claim 1, wherein comparing the first distribution and the second distribution includes comparing one or more of a mean, mode, and standard deviation of the first distribution to one or more of a mean, mode, and standard deviation of the second distribution.

5. The method of claim 1, wherein the first distribution and the second distribution are distributions of a weighted distance between a customer geographical location and commercial establishment geographical locations.

6. The method of claim 1, wherein the first distribution and the second distribution are distributions of a weighted drive time between a customer geographical location and commercial establishment geographical locations.

7. The method of claim 1, wherein modifying selection of entries in one or more of the training data set and the testing data set includes generating recommendations for improving selection of entries in one or more of the training data set and the testing data set.

8. The method of claim 1, wherein the training data set and the testing data set are selected from a customer information database.

9. The method of claim 1, further comprising comparing at least one of the first distribution and the second distribution to a distribution of a customer database.

10. The method of claim 1, wherein the first distribution and second distribution are frequency distributions of one of drive time and distance between a customer geographical location and one or more commercial establishment geographical locations.

11. The method of claim 9, wherein comparing at least one of the first distribution and the second distribution to a distribution of a customer database includes:

generating a composite data set from the training data set and the testing data set; and

generating a composite distribution from the composite data set.

12. The method of claim 1, wherein modifying selection of entries in one or more of the training data set and the testing data set includes changing one of a random selection algorithm and a seed value for a random selection algorithm.

13. The method of claim 1, further comprising training a predictive algorithm using at least one of the training

data set and the testing data set if the discrepancy is within a predetermined tolerance.

14. The method of claim 13, wherein the predictive algorithm is a discovery based data mining algorithm.

15. An apparatus for selecting data sets for use with a predictive algorithm, comprising:

a statistical engine; and

a comparison engine coupled to the statistical engine, wherein the statistical engine generates a first distribution of a training data set and a second distribution of a testing data set, the comparison engine compares the first distribution and the second distribution to identify a discrepancy between the first distribution and the second distribution and modifies selection of entries in one or more of the training data set and the testing data set based on the discrepancy between the first distribution and the second distribution.

16. The apparatus of claim 15, wherein the first distribution and the second distribution are distributions of drive time from a customer geographical location to a commercial establishment geographical location.

17. The apparatus of claim 15, wherein the first distribution and the second distribution are

distributions of distance between a customer geographical location and a commercial establishment geograhical location.

18. The apparatus of claim 15, wherein the comparison engine compares the first distribution and the second distribution by comparing one or more of a mean, mode, and standard deviation of the first distribution to one or more of a mean, mode, and standard deviation of the second distribution.

19. The apparatus of claim 15, wherein the first distribution and the second distribution are distributions of a weighted distance between a customer geographical location and commercial establishment geographical locations.

20. The apparatus of claim 15, wherein the first distribution and the second distribution are distributions of a weighted drive time between a customer geographical location and commercial establishment geographical locations.

21. The apparatus of claim 15, wherein the comparison engine modifies selection of entries in one or more of the training data set and the testing data set by generating recommendations for improving selection of entries in one or more of the training data set and the testing data set.

22. The apparatus of claim 15, further comprising a training data set/testing data set selection device that selects the training data set and the testing data set from a customer information database.

23. The apparatus of claim 15, wherein the comparison engine further compares at least one of the first distribution and the second distribution to a distribution of a customer database.

24. The apparatus of claim 15, wherein the first distribution and second distribution are frequency distributions of one of drive time and distance between a customer geographical location and one or more commercial establishment geographical locations.

25. The apparatus of claim 23, wherein the comparison engine compares at least one of the first distribution and the second distribution to a distribution of a customer database by:

generating a composite data set from the training data set and the testing data set; and

generating a composite distribution from the composite data set.

26. The apparatus of claim 15, wherein the comparison engine modifies selection of entries in one or more of the training data set and the testing data set by

changing one of a random selection algorithm and a seed value for a random selection algorithm.

27. The apparatus of claim 15, further comprising a predictive algorithm device, wherein the predictive algorithm device is trained using at least one of the training data set and the testing data set if the discrepancy is within a predetermined tolerance.

28. The apparatus of claim 27, wherein the predictive algorithm is a discovery based data mining algorithm.

29. A computer program product in a computer readable medium for selecting data sets for use with a predictive algorithm, comprising:

first instructions for generating a first distribution of a training data set;

second instructions for generating a second distribution of a testing data set;

third instructions for comparing the first distribution and the second distribution to identify a discrepancy between the first distribution and the second distribution; and

fourth instructions for modifying selection of entries in one or more of the training data set and the testing data set based on the discrepancy between the first distribution and the second distribution.

30. The computer program product of claim 29, wherein the first distribution and the second distribution are distributions of drive time from a customer geographical location to a commercial establishment geographical location.

31. The computer program product of claim 29, wherein the first distribution and the second distribution are distributions of distance between a customer geographical location and a commercial establishment geographical location.

32. The computer program product of claim 29, wherein the third instructions for comparing the first distribution and the second distribution include instructions for comparing one or more of a mean, mode, and standard deviation of the first distribution to one or more of a mean, mode, and standard deviation of the second distribution.

33. The computer program product of claim 29, wherein the first distribution and the second distribution are distributions of a weighted distance between a customer geographical location and commercial establishment geographical locations.

34. The computer program product of claim 29, wherein the first distribution and the second distribution are distributions of a weighted drive time between a customer

geographical location and commercial establishment geographical locations.

35. The computer program product of claim 29, wherein the fourth instructions for modifying selection of entries in one or more of the training data set and the testing data set include instructions for generating recommendations for improving selection of entries in one or more of the training data set and the testing data set.

36. The computer program product of claim 29, further comprising fifth instructions for comparing at least one of the first distribution and the second distribution to a distribution of a customer database.

37. The computer program product of claim 29, wherein the first distribution and second distribution are frequency distributions of one of drive time and distance between a customer geographical location and one or more commercial establishment geographical locations.

38. The method of claim 36, wherein the fifth instructions include:

instructions for generating a composite data set from the training data set and the testing data set; and

instructions for generating a composite distribution from the composite data set.

39. The computer program product of claim 29, wherein the fourth instructions for modifying selection of entries in one or more of the training data set and the testing data set include instructions for changing one of a random selection algorithm and a seed value for a random selection algorithm.

40. The computer program product of claim 29, further comprising fifth instructions for training a predictive algorithm using at least one of the training data set and the testing data set if the discrepancy is within a predetermined tolerance.